Applied Genome Research

Annotation

205048 & 205049

Boas Pucker

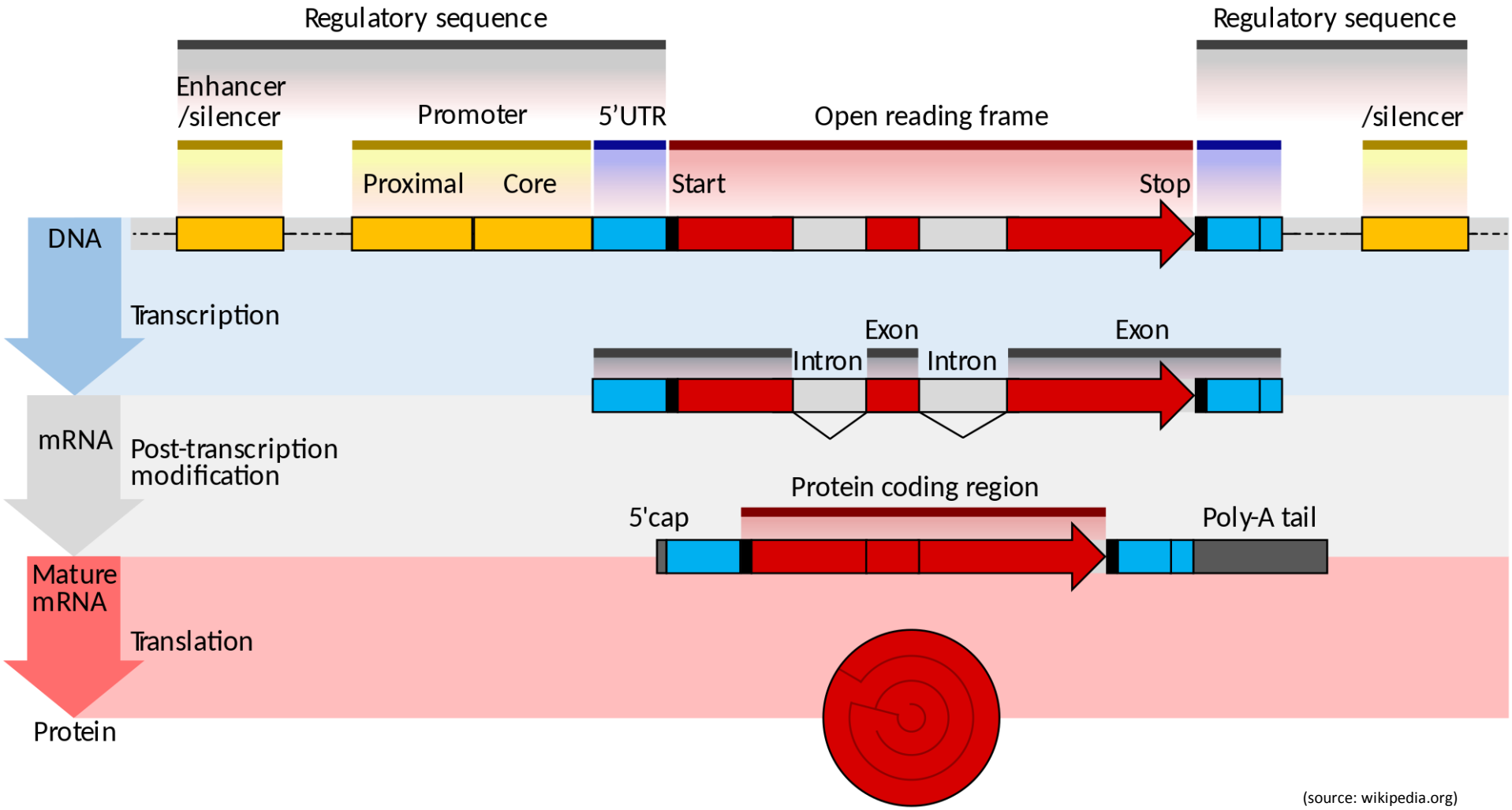
QUESTION

- What is a gene?

Prokaryotic gene structure

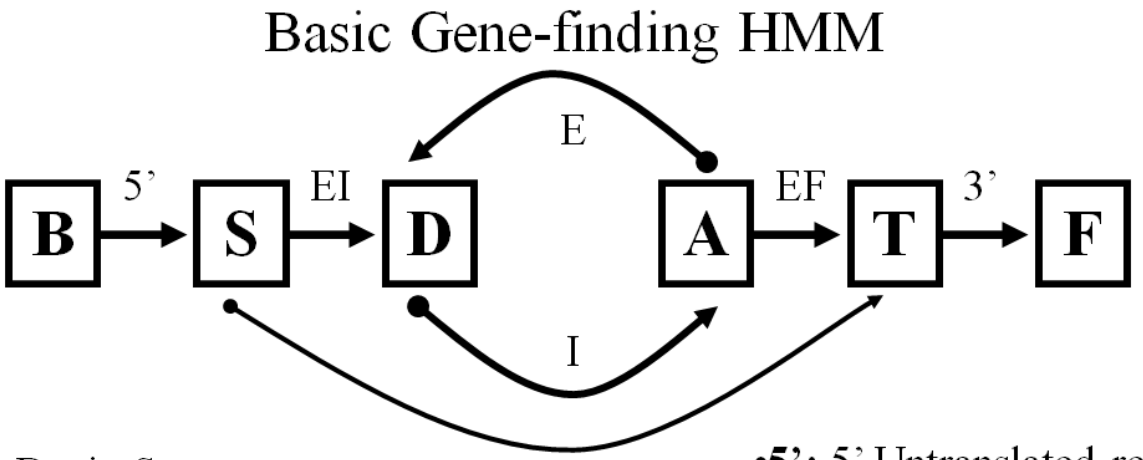
- Gene = sequence from start codon to stop codon (without any gaps)

Eukaryotic gene structure



(source: wikipedia.org)

Gene prediction - theory



- **B:** Begin Sequence
- **S:** Start translation
- **D:** Donor splice site
- **A:** Acceptor splice site
- **T:** Stop translation
- **F:** End sequence

- **5':** 5' Untranslated region
- **EI:** Initial exon
- **ES:** Single exon
- **E:** Exon
- **I:** Intron
- **EF:** Final exon
- **3':** 3' untranslated region

(source: blogspot.com)

Gene prediction - different modi

- *ab initio* = gene prediction without any additional information
- hint guided = RNA-Seq data or similar information are used to support prediction
- Reference-based = map annotated sequences from a reference to a newly assembled sequence

Hint-based gene prediction

- RNA-Seq reads or known transcript sequences are mapped

- Due to a lack of RNA-Seq data for Nd-1 gene prediction will be computed *ab initio*.

Gene prediction - AUGUSTUS

$ augustus-3.2\

--species=<SPECIES> … name of parameter set to use

--gff3=on … output format is GFF3

--codingseq=on … encoded protein sequences are written in output files

<FASTA_FILE_NAME>

> …. write output into file (instead of printing it to screen)

<outputfile>

Running *ab initio* gene prediction.

Gene prediction - AUGUSTUS

```
$ getAnnoFasta.pl\
 --seqfile=<assembly_file>\
 <GFF3_file>
```

Extraction of sequences of different predicted features.

EXERCISE

- Run AUGUSTUS on SOAPdenovo2 contigs!
- Run AUGUSTUS on SOAPdenovo2 scaffolds!
- Run AUGUSTUS on SSPACE scaffolds!
- Analyze the differences!

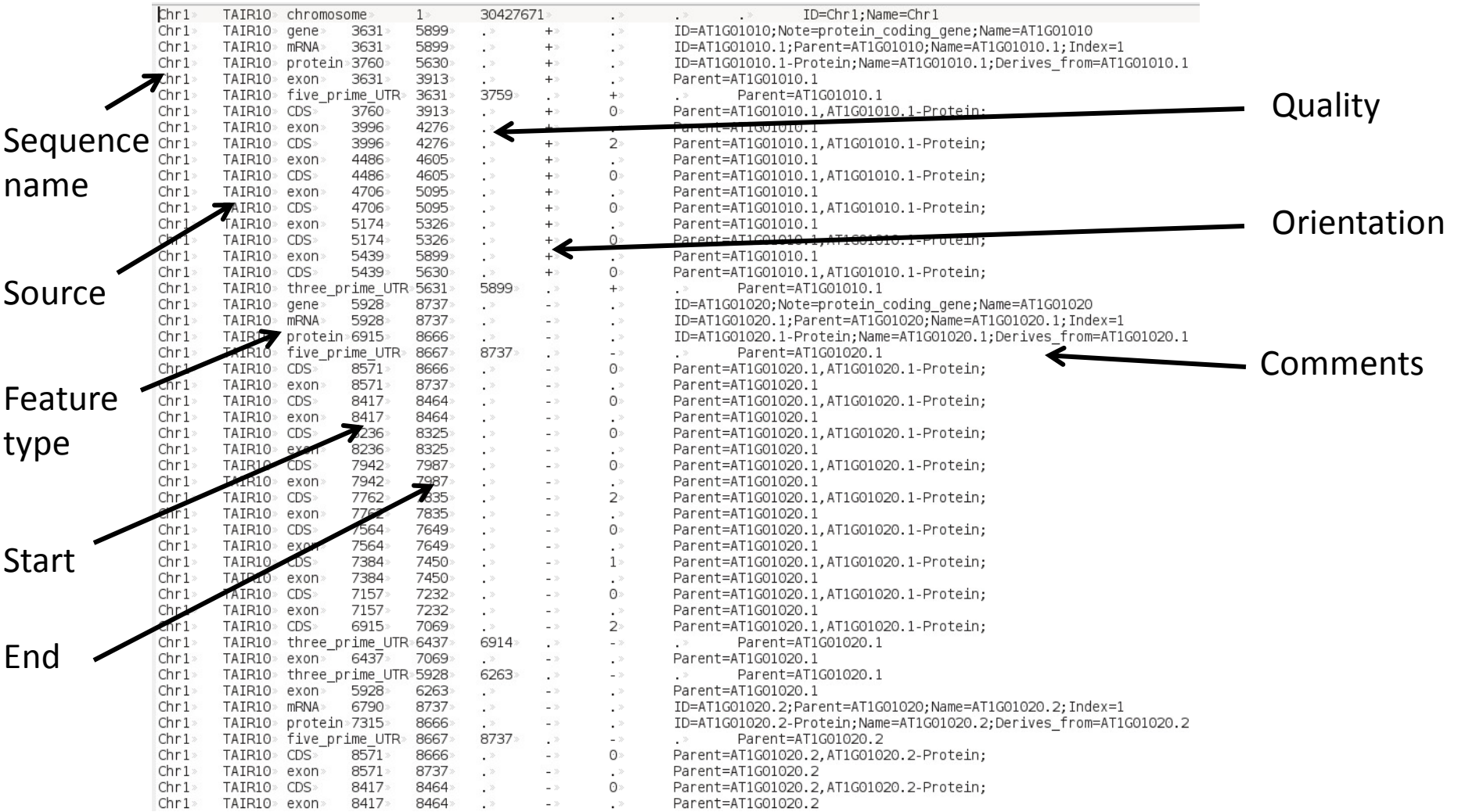
- Find ‘getAnnoFasta.pl’ (online), download it, and use it!

‘Fragmented genes’

Short read assembly:

Reality:

GFF = Generic Feature Format

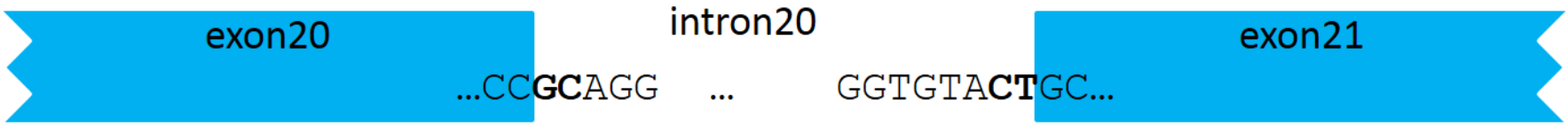


Gene prediction – reference-based

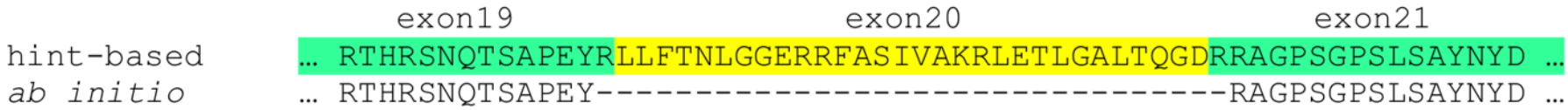
- *Arabidopsis thaliana* Col-0 reference sequence is very well annotated
- Mapping of Col-0 sequences to Nd-1 assembly might be useful
- Annotation data are available via TAIR10 and Araport11

Example: non-canonical splice sites

a



b



(source: Pucker et al., 2017)

- Only canonical splice sites (GT-AG) can be predicted *ab initio*
- Hint-based prediction of gene structures is more accurate

QUESTION/EXERCISE

- Have a look at TAIR10 and Araport11!
- How many protein-coding genes are annotated in Araport11?
- How can you find publications describing certain genes?
- Are there websites for other model organisms?
- List some databases for (model) organisms!

BLAST – sequence comparison

- Very efficient tool for sequence comparison
- Very highly cited publication by Altschul *et al.*, 1990 (indicates importance)
- Comparison is based on matching words (seeds) which are than extended to final alignments
- NCBI offers web-based service
- Command line version of BLAST is much more efficient (for large data sets)
- Submitting data to the NCBI is not always allowed (e.g. clinical data)

BLAST – command line

$makeblastdb –in <fasta_file> -out <some_name> -dbtype ‘nucl’

$ blastn\ …. There are different verions of BLAST (n/p/x…)

-query <query_file> … file with sequences to search for

-subject <subject_file> … file with sequences to search in

(-db <some_name>)

-out <output_file> … file for results

-outfmt 6 … set output format to table

-evalue 0.001 … set e-value cutoff

(-num_threads 4 … set number of threads to use for search)

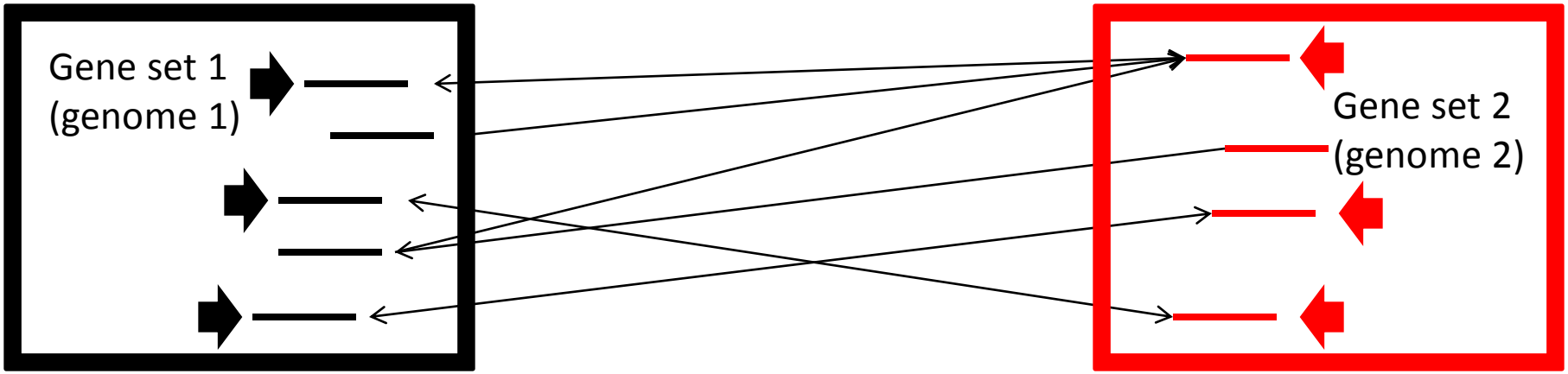
EXERCISE

- Run BLASTn search of Col-0 exons vs. SOAPdenovo2 contigs of Nd-1!
- How many hits?
- What is the next step?

BLAST – annotation transfer

- High sequence similarity indicates a common function of two genes/proteins
- Identification of Reciprocal Best Hits (RBHs) can be used to transfer functional annotations
  - Two data sets:
    - SeqA, SeqB, SeqC, ….
    - Seq1, Seq2, Seq3, …
  - Best hit of SeqA is Seq2 and best hit of Seq2 is SeqA => RBH

Gene set comparison

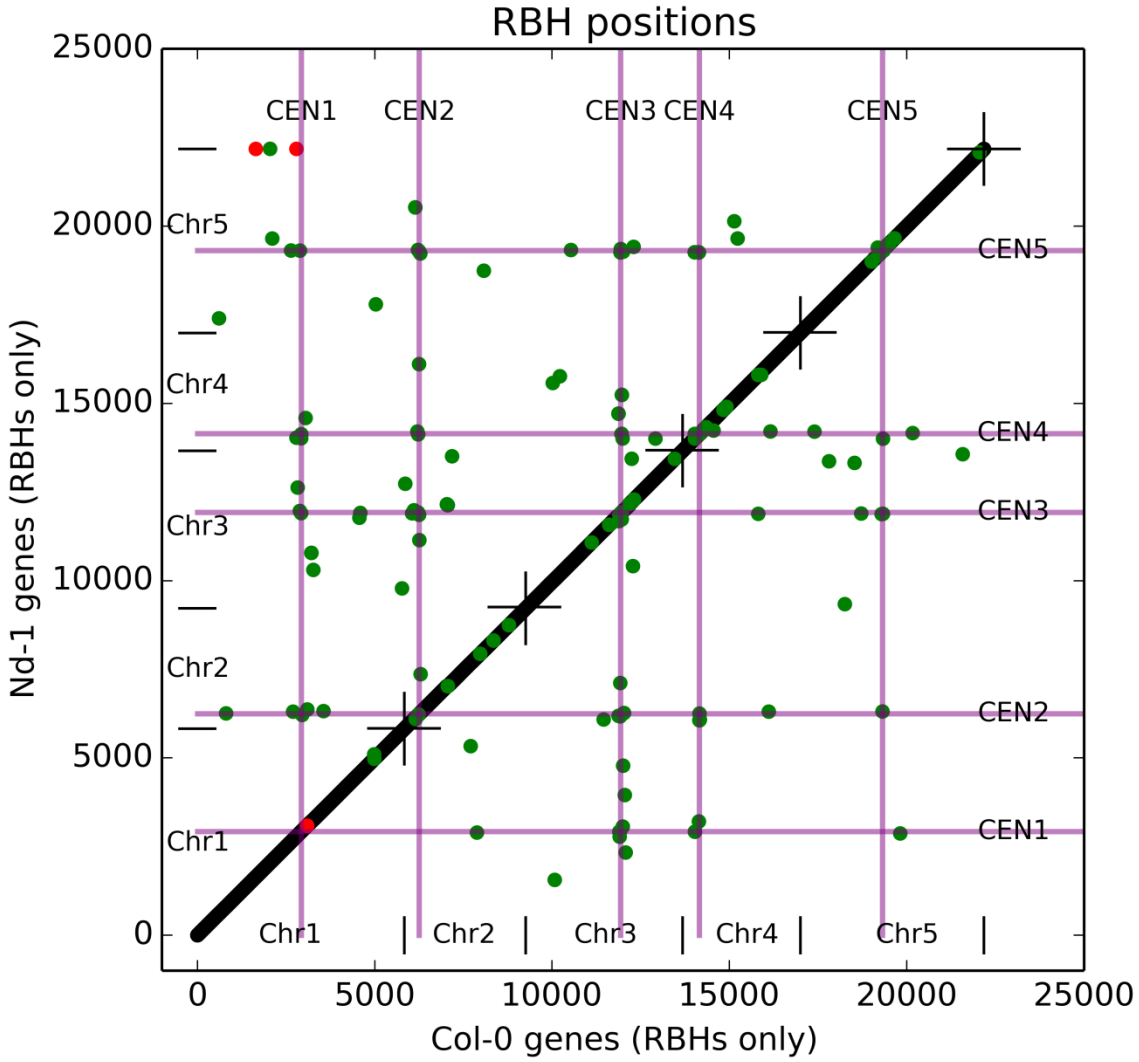


- RBH = Reciprocal Best (BLAST) Hit
- Comparison can be done on DNA or peptide sequence level
- Broad range of applications!

EXERCISE

- Identify RBHs between Col-0 and Nd-1 on peptide sequence level via identify_RBHs.py!
- How many RBHs are there?
- Map TAIR10 annotation to the identified RBHs via map_annotation.py!
- Select one gene and collect additional (functional) information!
- Collect information about all these candidate genes and find a systematic pattern!

Application of RBHs



(source: Pucker et al., 2016)